- 1 -

A LINK GENERATION SYSTEM

FIELD OF THE INVENTION

The present invention relates to a link generation system and process for generating links

5   for a network server or structured data set such as a web site.

BACKGROUND

The ever increasing amount of information available on the Internet can make it extremely
difficult to locate information relevant to a topic of interest. In the case of information

10   available on the world-wide web, search engines have been developed for generating lists
of hypertext markup language (HTML) documents or web pages matching one or more
search terms supplied by a user. These lists of pages are generated from inverted indices
generated by analysing the content of individual web pages. These web pages are retrieved
by software modules known as spiders or web-crawling agents that crawl the web, using

15   the hypertext transfer protocol (HTTP) to retrieve individual web pages, analyse content of
those pages, and generate indices. This may involve identifying hyperlinks to other web
pages, retrieving those linked pages, and analysing their content. Spiders can be used to
generate indices for the world-wide web itself, or can be restricted to one or more specified
web sites.

20

A web site can be viewed as a directed graph or digraph, with the servable content forming
the nodes in the graph and directed links between the nodes corresponding to hypertext
links within web pages of the site. A spider begins at one of the nodes in a web site, and
then follows the links from that node to other nodes, and so on. The spider can perform

25   whatever processing is desired for the nodes as it encounters them. In the case of a search
engine spider, this involves indexing node content, but other spider types can be used to
perform other tasks such as checking for broken hyperlinks or spell checking documents.

Unfortunately, not all web sites are completely connected - many have pages that are not directly connected to the rest of the web site through a hypertext link. In such a disconnected web site, a spider is unable to visit all of the nodes of the web site. This problem is especially pronounced in sites whose web pages include dynamic content. In
5   the case of an indexing spider, a significant proportion of a site's content may not be accessible by a corresponding search engine. As more web sites convert their content from pre-existing, static web pages to more flexible and easier to maintain web pages including dynamically generated content, this problem will become even more significant.

10 Lack of full connectedness in a web site is also a potential problem for web site administrators who are trying to track their site's content. Without a completely connected graph of the site, it can be a difficult task to find all of the site content. For large sites with many content contributors, this task can become almost impossible.

15 Content that is not indexed by search engines has been referred to as 'the invisible web,' because it is not generally visible. It has even been suggested that the majority of information available on the web is invisible. Because invisible content is inaccessible to search engines, it decreases the visibility of web sites with invisible content, and degrades the usefulness of the web in general by making such content difficult to find.
20

It is desired, therefore, to provide a link generation system and process that alleviate one or more of the above difficulties, or at least to provide a useful alternative to existing link generation systems and processes.

SUMMARY OF THE INVENTION

25 In accordance with the present invention there is provided a link generation process executed by a computer system, including:

processing data files of a network server to identify servable data; and

generating links to said servable data to allow said servable content to be accessed using said links.

- 3 -

The present invention also provides a link generation process executed by a computer system, including:

    generating one or more data nodes representing servable data of a network server;

5 and

    generating one or more links for an indexing agent on the basis of said data nodes for retrieving said servable data.

The present invention also provides a link generation process executed by a computer

10 system, including generating at least one encoded link for retrieving dynamic content data of a hierarchical data set in response to selecting said at least one encoded link.

The present invention also provides a link generation process, including:

    generating links for dynamic content of a network site;

15   receiving requests from an indexing agent for content of said site; and

    responding to said requests with said links and said dynamic content corresponding thereto for indexing.

The present invention also provides a link generation system, including:

20   a content discovery module for processing data files of a network server to identify servable data; and

    a link generator for generating links to said servable data to allow said servable content to be accessed using said links.

25 The present invention also provides a link generation system, including:

    one or more content discovery modules for processing data files of respective network servers to identify servable data; and

    a link generator for generating links to said servable data to allow said servable content to be accessed using said links.

30

- 26 -

CLAIMS:

1. A link generation process executed by a computer system, including:

   processing data files of a network server to identify servable data; and

5 generating links to said servable data to allow said servable content to be accessed using said links.

2. A process as claimed in claim 1, wherein said generating includes generating encoded links for accessing dynamically generated data of said network server, said encoded

10 links being in a form suitable for an indexing agent.

3. A process as claimed in claim 2, wherein said encoded links are URI-encoded.

4. A process as claimed in claim 2, wherein said processing includes processing a

15 database of said network server to determine query data for retrieving servable data from said database.

5. A process as claimed in claim 4, including processing scripts of said web site to determine request data for retrieving said servable data; wherein said encoded link is

20 generated on the basis of said request data and said query data.

6. A process as claimed in claim 5, wherein said step of processing scripts includes processing said scripts to determine access data for accessing said database.

25 7. A process as claimed in claim 2, including receiving a request generated in response to selecting one of said encoded links, translating said request, and forwarding the translated request to said network server to access corresponding dynamically generated data of said network server.

8. A process as claimed in claim 7, wherein said translated request is an HTTP GET request.

9. A process as claimed in claim 7, wherein said translated request is an HTTP POST request.

10. A process as claimed in claim 1, including sending said links to a remote indexing agent to allow said servable data to be indexed.

11. A process as claimed in claim 1, including sending said encoded link to a remote system using one of HTTP PUT, HTTP POST, FTP, and SMTP.

12. A process as claimed in claim 1, wherein all servable data of said network server can be accessed via selection of any one of said links.

13. A link generation process executed by a computer system, including:

generating one or more data nodes representing servable data of a network server; and

generating one or more links for an indexing agent on the basis of said data nodes for retrieving said servable data.

14. A link generation process as claimed in claim 13, wherein said data nodes represent data for which no link exists in said data set.

15. A link generation process as claimed in claim 13, wherein said data corresponds to static data of said data set.

16. A link generation process as claimed in claim 13, wherein said data corresponds to dynamic data of said data set.

17. A link generation process as claimed in claim 13, wherein said nodes are organised in a hierarchical structure representing a hierarchical structure of said data set.

18. A link generation process executed by a computer system, including generating at least one encoded link for retrieving dynamic content data of a hierarchical data set in response to selecting said at least one encoded link.

19. A link generation process as claimed in claim 18, including generating a list of links to content data of at least one node of said hierarchical data set, said links including said at least one encoded link.

20. A link generation process as claimed in claim 19, said links include links to all available data of said hierarchical data set.

21. A link generation process as claimed in claim 19, wherein said links include one or more direct links to content data of said hierarchical data set.

22. A link generation process as claimed in claim 19, wherein said links include one or more indirect links to content data of said hierarchical data set.

23. A link generation process as claimed in claim 19, wherein said links include at least one of a direct and an indirect link to content data of said node.

24. A link generation process as claimed in claim 19, wherein said list of links corresponds to a node of said hierarchical data set.

25. A link generation process as claimed in claim 19, wherein said hierarchical data set includes at least one web site.

26. A link generation process as claimed in claim 18, wherein said at least one encoded link includes at least one encoded POST query.

27. A link generation process as claimed in claim 18, wherein said at least one encoded
5   link includes at least one encoded GET query.

28. A link generation process, including:

    generating links for dynamic content of a network site;

    receiving requests from an indexing agent for content of said site; and

10   responding to said requests with said links and said dynamic content corresponding thereto for indexing.

29. A process as claimed in claim 1, wherein said links are generated as one or more of hyperlinks, XML elements, and text.

15

30. A link generation system having components for executing the steps of any one of claims 1 to 29.

31. Link generation software having program code for executing the steps of any one of
20   claims 1 to 29.

32. A computer readable storage medium having stored thereon program code for executing the steps of any one of claims 1 to 29.

25   33. A link generation system, including:

    a content discovery module for processing data files of a network server to identify servable data; and

    a link generator for generating links to said servable data to allow said servable content to be accessed using said links.

30

- 30 -

34. A link generation system as claimed in claim 33, wherein said link generator is adapted to generate encoded links for accessing dynamically generated data of said network server, said encoded links being in a form suitable for an indexing agent.

5   35. A link generation system as claimed in claim 34, wherein said link generator is adapted to process a database of said network server to determine query data for retrieving servable data from said database.

36. A link generation system as claimed in claim 35, including a proxy server for receiving
10   a request generated in response to selecting one of said encoded links, translating said request, and forwarding the translated request to said network server to access corresponding dynamically generated data of said network server.

37. A link generation system, including:
15   one or more content discovery modules for processing data files of respective network servers to identify servable data; and
    a link generator for generating links to said servable data to allow said servable content to be accessed using said links.

20